DOCUMENT RESUME

ED 166 252 TM 008 292

AUTHOR Kennedy, Dale
TITLE The Nebraska Blueprint for School Assessment Programs. Volume 3: Technical Information.
INSTITUTION Nebraska State Dept. of Education, Lincoln.
PUB DATE 78
NOTE 77p.; For related documents, see TM 008 290 and 291

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS *Data Analysis; *Educational Assessment; *Educational Testing; Elementary Secondary Education; Glossaries; Guidelines; Resource Materials; Testing Programs; Test Reliability; *Tests; *Test Selection; Test Validity

ABSTRACT

This is the third of three volumes designed to cover the problems and issues of developing and managing a formal assessment program. They also offer guidelines and illustrations of several of the more special aspects of assessment programs, such as administering tests or documenting their results. This volume contains summaries of technical information. It discusses test reliability and validity. It lists the addresses of several test publishers and provides a classified list of 90 tests or test batteries including achievement tests, diagnostic tests, interest tests, mental ability tests, multi-aptitude tests, personality tests, reading readiness tests, reading tests, and other tests. It lists a variety of sources of information, including professional associations and publications. It provides computing guides for common elementary statistical procedures and a glossary of 67 statistical or measurement terms. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED166252

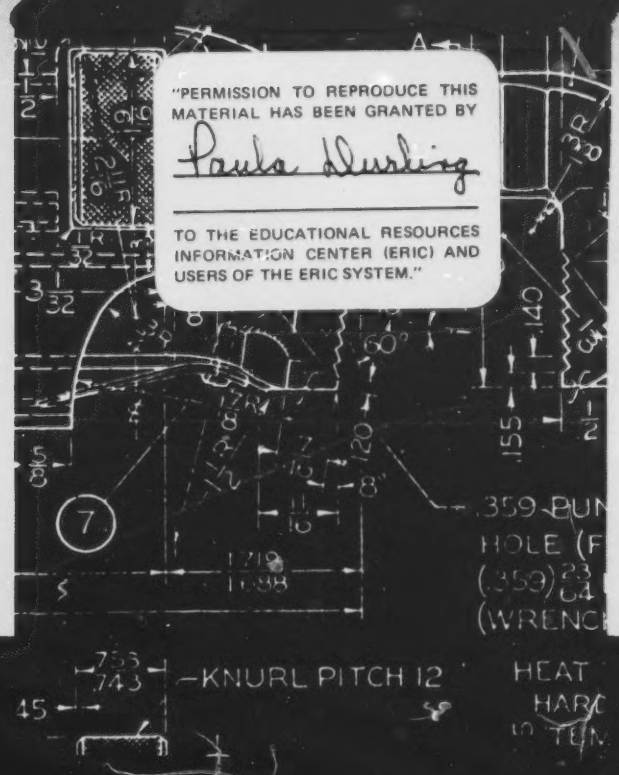
U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Paula Hurling

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."



The Nebraska

BLUE PRINT

For School Assessment Programs

Volume 3: Technical Information

Issued by the Nebraska State Department of Education

M. Anne Campbell, Commissioner of Education

301 Centennial Mall South
P. O. Box 94987
Lincoln, Nebraska 68509

1978

TABLE OF CONTENTS

Volume III: Technical Information

1. Reliability 1
   - Types 1
     - Stability 1
     - Equivalence 1
     - Internal Consistency 2
2. Validity 3
   - Types 3
     - Predictive Validity 3
     - Concurrent Validity 4
     - Content and Construct Validity 4
3. Test Tables 6
   - Test Publishers T1
   - Achievement Tests T2
   - Diagnostic Tests T5
   - Interests Tests T8
   - Mental Ability Tests T11
   - Multi-Aptitude Tests T13
   - Personality Tests T14
   - Reading Readiness Tests T15
   - Reading Tests T16
   - Other Tests T19
4. Information Sources 25
   - Professional Associations 25
   - Newsletters, Special Publications, Professional Journals 25
   - Useful References 26
   - External Testing Programs 27
   - Test Instruments Library 27
5. Computing Guides 30
   1. Determining Percentile Equivalents 31
   2. How to Compute Local Norms 33
   3. Calculation of Standard Scores 40
   4. Relations Between Standard Scores and Percentile Scores, When Raw Scores are Normally Distributed 41
   5. Methods for Estimating Expected Scores 42
   6. Use of Spearman-Brown Formula 45
   7. Internal Consistency Formulas 46
   8. Stanine Computation for Sample Mathematics Test Scores 47
   9. Reliability of Differences Between Part Scores 48
   10. The Standard Error of Measurement and Confidence Bands 49
6. Glossary 51

COMMISSIONER'S FOREWORD

Assessment, in ordinary words, is the effort devoted to finding out if students are learning what we think we are teaching. Like instruction, it is a complex undertaking, both in conception and in practice. Both aspects, instruction and assessment, are essential to the process of education and both require careful and thorough development to be of best effect. The Department's Blueprint for School Assessment Programs covers the problems and issues of developing and managing an assessment program. It also offers guidelines and illustrations of several of the more special aspects of assessment programs, such as administering tests or documenting their results. The Blueprint's coverage is comprehensive and will have value for a variety of audiences that may be involved in assessment programming.

It is anticipated that the Blueprint will prove a useful reference in planning, developing, evaluating, reviewing, and managing local assessment programs.

M. Anne Campbell
Commissioner of Education

AUTHOR'S FOREWORD

The Blueprint is intended to be a repository of theoretical and practical information on student assessment and the problems of assessment programming. It has the qualities each of a textbook, a handbook, and a reference manual. It is possible for it to be read and used at several different levels. It could satisfy readers having general interest in assessment theory and technology, but it should also prove to be of repeated value for readers who are actually involved in assessment programming, such as the planning committee members, examiners, and the persons charged with the school or district's assessment responsibilities. It may be scanned for general information or it may be studied for precise and technical information. It can serve to generally refresh and improve one's overall understanding of the issues and problems of assessment, or it can be employed as a guidelines and procedures manual for the actual planning and operation of an assessment program.

Those persons who are actually involved in the planning and operation of a school or district assessment should seriously consider acquiring and reading Standards for Educational and Psychological Tests, published by the American Psychological Association, Guide For School Testing Programs, published by the National Council on Measurement in Education, and Essentials of Psychological Testing, by Lee J. Cronbach, published by Harper and Row. There are other good books on psychological testing besides Cronbach's. The Essentials, however, exhibits the unique combination of well explained theory along with thorough discussions of practical concerns. Cronbach also writes in a scientific style

that is both exacting and conservative in statement. These are qualities which the Blueprint also attempts to achieve, though in a more synoptic form. The Blueprint, together with these three references, should soundly provide a thorough and comprehensive knowledge of assessment theory and practice.

Dale Kennedy

PREFACE

Human children are growing, learning, maturing, thriving organisms. The ingredients of their perceptions and cognitions, of their learning, far exceed those of other living things found upon earth, for children's growth must take place on several levels simultaneously: physical, intellectual and emotional. The purpose of education, then, is to shape and coax this growth (if not to stimulate it outright). Education, when harnessed to growth, leads it to its most mature levels in a way that ideally enhances further self education. An education "completed" ideally leaves the student open for a multitude of future circumstances.

Learning takes place as a function of three things: 1) past learning, in which one can include native ability because its effects are in fact manifested by the past learning; 2) the occasion upon which the learning takes place, i.e., the learning opportunity to the teacher and the learning experience to the student; and 3) the consequences contingent upon the learning behavior itself. Learning occurs if the behavior of the students takes a new shape during and after the occasion.

The change is lasting or strong, occurring again on appropriate future occasions or affecting future learning, in proportion to the reinforcement of the learning on the occasion of first change. But the student must first perceive the change and the consequences thereof. If the consequences are natural, as in the discovery of the world of literature through reading, the student must value these new possibilities. Teachers must also perceive the change if their actions toward the students are to be consistent, timely, and in pace with the students'

rate of growth. This reading of student behavior and ability is assessment.

Most assessment is informal, taking place each day, any time the teacher and student interact. Most of the student's self-assessment also takes place on informal occasions. The assessment function becomes formal when the student is asked to perform on occasions which would not normally be encountered as a daily routine, designed to reveal something of the student's ability. These occasions of formal assessment are the subject matter of the Blueprint.

The Blueprint is published in three separate volumes. Volume I, Purposes and Categories of Measurement, deals with the regulations governing school assessment programs, the general purposes for which assessment information is collected, and the kinds of measuring instruments used for such collection. This is the most conceptually involved volume, but it does provide a comprehensive view of existing measurement theory and measurement technology with respect to educational settings.

Volume II, Processes and Procedures, deals with planning of assessment programs, the program's standing documentation, and problems of administering, scoring, record-keeping and reporting, referred to collectively in the Blueprint as the logistics of management. The information provided in this volume balances the conceptual content of Volume I with detailed exposition on practical considerations. Volumes I and II are intended to complement each other.

Volume III, Technical Information, is nothing more than a repository of useful information of a technical or abbreviated nature. It has sections concerning reliability and validity, addresses of test publishers,

lists of references, computing guides for various statistics, test tables and a glossary. The information in this volume is also of a practical nature, even more so than Volume II.

There are four central assumptions underlying all volumes of the Blueprint.

First, a district's assessment program already exists and stands in a working relationship with the operations of instructional delivery. The primary object of the Blueprint is to provide the concepts and procedures that will bring the documentation of the program to a comprehensive, organized form, if it does not already exist in such a systematic fashion.

Second, the assessment of individuals and groups must be fair to those individuals and groups. If a test discriminates against an individual on the basis of background or physical handicap, the results will not represent true information on the individual's abilities or achievement. If used to make decisions about the individual's future education, such unfairly obtained data would be misinformation causing mistakes in judgment that may result in the miseducation of the individual. The fairness stricture also applies to mainstream groups. Here, if the administrations are poorly controlled, the margin of error is enlarged and the test becomes an inaccurate measure mostly containing misinformation. Assessments must be exactly administered if they are to be fair. Miseducation is the consequence of mistaken judgment.

The possibilities of miseducation are indeed some of the central reasons for establishing and maintaining an assessment program. For example, it is always desirable to fit instruction to known strengths

and weaknesses of the student, or groups of students. The educative process would be expected to reinforce the strengths, improve the weaknesses, and in its combination of effects stimulate an optimum amount of learning. But just as surely as instruction can promote optimum growth, it can drop a ceiling on learning. Instruction can mismatch, as well as match, the strengths and weaknesses of students. And in so doing, it fails to promote growth where potential is great or expects more of the student than the student is truly capable of. A mismatch guarantees a growth far less than what is optimum and, in effect, blocks such growth. Miseducation is simply the diminishment of learning through an inappropriately applied (indeed, an alleged) educative process. It is always a danger. Good assessment information should reduce and limit occurrences of this plight.

Third, grade promotion and graduation sanctions, contingent upon assessment information, are not necessary to effective and useful assessment programs. They are essentially a special credit system extended beyond the usual credits for attendance and passing courses. Where they are instituted, they became a defining feature of a whole assessment program, yet a program has several good uses without them. The logic and rationale of grade promotion and graduation requirements are not topics of the Blueprint. The subjects that are discussed, however, would give a soundness to the assessment information, so that if used for promotion, the information would at least be validly obtained.

Fourth, the readers of this manual and all those involved in assessment, either as tester, subject, or interpreter, should understand and appreciate the scientific nature of educational assessment. An interpre-

tation made from assessment information is a work of scientific reasoning. Both the proponents and the critics of scientific method as applied to education tend to regard the mechanical qualities of the methods of science as its only virtue, or weakness, as the case may be. This neglects the deliberateness and natural conservatism of scientific method. It also misses the cultivation of balanced perspective. Whether very mechanically obtained, or more informally so, assessment information must: first, be the right information needed; second, be accurate; and third, be viewed in relation to other assessment information on the same individual or group. If the information does not fit the questions asked, or if it is error-ridden, then it is misinformation and, if interpreted, would foster mistaken judgments and lead to the probable miseducation of the students. It requires an understanding of scientific method to recognize when and under what conditions the information gathered for the purposes of assessment proves to be misinformation, data which can be harmful to the future educational prospects of students.

Above all else, this scientific perspective must be grounded in a concrete understanding of the conditions under which the data are actually produced. It is the children who face the test and perform as best they can that day. The temperature of the room, the lighting, their breakfast or lack of breakfast, and many other daily, mundane facts, all conspire to affect the children's performance. The mechanical rigors of administration are intended to even out the disadvantages and to heighten the individual's performance. A testing procedure cannot be too precise when these are its objects.

TECHNICAL INFORMATION

The sections that follow provide a background of technical information on some of the key concepts of the previous section of the Blueprint. In particular, there are more exact treatments of reliability and validity, a glossary of terms, computing guides, and resource lists and tables. This information supplements the general outlines of assessment programs and yet concentrates on practical needs.

Section	Contents
Section 1	Reliability
Section 2	Validity
Section 3	Test Tables
Section 4	Information Sources
Section 5	Computing Guides
Section 6	Glossary

Section 1

RELIABILITY

Reliability refers to the consistency of an instrument over a series of measurements. Reliability coefficients are denoted symbolically as "r" with double subscripts, r_{11}, for instance. The value of a reliability coefficient expresses the proportion of nonerror or true variance to the total test's variance. It depends upon test length and the spread of the group of raw scores for which it is computed; the longer a test and the greater the diversity of obtained scores, the more likely the test will prove reliable. It is related to validity by the following rule, "The validity coefficient cannot exceed the square root of the reliability coefficient."

Types[1]

Stability - Coefficients of stability are obtained by correlating one set of test results with retest results obtained from the same individuals under similar conditions. The time elapsed between test and retest will vary. In the case of achievement tests, too long a time between test and retest allows retest results to be influenced by future learning. But in the case of diagnostic tests, there is stability over longer periods of time, several years at least with intelligence tests. Otherwise, the diagnostic value would be temporary.

Equivalence - Coefficients of equivalence are obtained by correlating the performance of a group of individuals on equivalent forms of the same test at the same time of administration. Where time elapses between the administration and times of the equivalent forms, the coefficient

---

[1]Condensed from Essentials, pp. 126-144.

crosses equivalence and stability.

Internal Consistency - Where equivalent forms do not exist or where retesting is impractical, coefficients of internal consistency can be obtained.

In the split half method, a test is administered in the usual way, but the results are scored in two independent halves. The halves can be obtained by randomly dividing the items into two parts. The coefficient can be computed in either of two ways.

i. By the formula

$$r_{tt} = 2\left(1 - \frac{S_a^2 + S_b^2}{S_t^2}\right)$$

where r_{tt} = the reliability coefficient

S_a^2 and S_b^2 = the variance of each split half

S_t^2 = the total test variance

ii. By correlating the split halves, producing an equivalence coefficient denoted as $r_{\frac{1}{2}\frac{1}{2}}$ and using this value in the Spearman-Brown Formula in the case of estimating the reliability of a test of twice the given length. (See Computing Guide 6.)

There are three widely-used formulas for computing a coefficient of internal consistency without involving the splitting procedure: the Kuder-Richardson estimates; K-R 20 and K-R 21; and coefficient alpha.

These formulas consider the average correlation obtained from all possible split-half reliability coefficients. (See Computing Guide 7.)

There are four principal types of validity: predictive; concurrent; content; and construct. One or more will be applicable for any test. The degree of validity of a test represents the degree to which it truly performs the work it is intended to do, e.g., it truly covers the content, it truly predicts success in college.

Predictive Validity

Diagnostic information is used to assess an individual's strengths and weaknesses, his interests and capabilities. Sooner or later, each such assessment will face a situation confirming or disconfirming the original assessment, e.g., college bound students will have undergraduate G.P.A.s, elementary students will have to learn more things. Where the future performance of interest can itself be measured, this measure stands as a criterion against which the diagnostic information can be validated. The validity coefficient is simply the correlation between scores on the diagnostic test and the criterion measure. The results of predictive validation studies should be reported in tests manuals accompanying the specimen set.

A district or school can also perform its own validation studies of this sort. As part of its accountability function, an assessment program will have a systematic record on individual passage and progress through the school's learning opportunities. This data represents the criterion measures required for validating any diagnostic information. Provided that the school is satisfied with the content and construct validity of its criterion measures, it should expect its diagnostic instruments to have reasonably good, middle range correlation with them.

Section 2

VALIDITY

If, for instance, a reading diagnostic battery shows no relationship to reading achievement a year or so later, it is probable that the battery is producing misleading information.

Concurrent Validity

Predictive validity is founded on the correlation between immediate test results and a criterion measure of future performance measured at a later date. Where the criterion measure is administered at the same time as the diagnostic instrument, or is available and is collected at the same time, the correlations are coefficients of concurrent validity. These correlations are usually computed against an instrument that has already been validated and that purports to measure the same things as the instrument in question. One can also compare diagnostic results with past achievement. In any case, concurrent validation should only be temporary evidence on criterion-related validity. The true test is predictive validity, the correlation with future performance.

Content and Construct Validity

The first two types of validity, predictive and concurrent, concern the relationship of the test with some criterion external to the test. The next two types of validity, content and construct, concern the relationship of the test and the thing it purports to measure. Content validation typically involves matching the test's items with the test's stated content specification and asking the question, "Does the item represent the content it is intended to measure?" When an instrument is used as a school criterion measure, it is also necessary to ask,

"Does this item represent content that was taught in the local curriculum?" Procedures for local content validation are discussed in Volume I, "Varieties of Tests and Assessment Techniques" (pg. 25 to 27). Content validation is strictly a logical process.

When the things measured can be interpreted in terms of psychological concepts and constructs, instead of logical divisions of subject matter, the type of validity involved is construct validity. Skills tests, general and special intelligence tests, and personality measures involve this kind of validity. In its simplest forms, construct validation follows mainly from an examination of items answering, for example, a question such as, "Does the item represent a valid instance of a skill application?" Yet because the measure in question is interpreted as a psychological capability, the measure should bear certain predictable relationships to external criterion measures. Providing such evidence frequently involves experimental research design and consequently most construct validation is reserved for experimental psychologists. The local district, however, can assure itself of at least the simple form of construct as content validation when the thing measured appears to be more of an individual trait or behavioral capability rather than a content area in the body of human knowledge.

SECTION 3

Test Tables*

*Note: The following is not a complete listing of available tests. They only serve as examples of subject areas in which tests have been developed.

A REPRESENTATIVE LIST OF TEST PUBLISHERS

Counselors often need to communicate with test publishers, and the addresses of some of the companies are listed below for your convenience.

A. Bobbs-Merrill Company, Inc.
4300 East 62nd Street
Indianapolis, Indiana 46268

B. Bureau of Publications
Teachers College Press
502 West 121st Street
New York, New York 10027

C. California Test Bureau/McGraw Hill
Del Monte Research Park
Monterey Park, California 93940

D. Educational Testing Service
Rosedale Road
Princeton, New Jersey 08540

E. Ginn and Company
450 West Algonquin Road
Arlington Heights, Illinois 60005

F. Harcourt, Brace & Jovanovich, Inc.
757 Third Avenue
New York, New York 10017

G. Houghton-Mifflin Company
110 Tremont Street
Boston, Massachusetts 02107

H. Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, Illinois 60106

I. Psychological Corporation
755 Caldwell Avenue
Chicago, Illinois 60648

J. Science Research Associates
259 East Erie Street
Chicago, Illinois 60611

K. Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, California 90025

L. Committee on Diagnostic
Reading Tests, Inc.
Mountain Home
North Carolina 28758

M. Lyons and Carnahan
407 East 25th Street
Chicago, Illinois 60616

N. Personnel Research Assoc., Inc.
701 Metropolitan Building
1407 Main Street
Dallas, Texas 75202

ACHIEVEMENT TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
CALIFORNIA ACHIEVEMENT TEST					
11 or 12 scores: Reading (vocabulary, comprehension, total), Mathematics (computation, concepts and problems, total), Language (auding (level I only) Mechanics, usage and structure, total, spelling), total.	1.5-2 2-4 4-6 6-9 9-12	1970	Level 1 - 114 min. Level 2 - 123 min. Level 3 - 152 min. Level 4 - 198 min. Level 5 - 153 min.	C	CompuScan Digitek IBM 1230
IOWA TESTS OF BASIC SKILLS					
Primary: Listening, word analysis, reading comprehension (3 scores), vocabulary, spelling, capitalization, punctuation, usage, maps, graphs and tables, references, mathematics, concepts, mathematics problems. Grades 3-8: Vocabulary, reading comprehension, language skills (4 scores), work study skills (3 scores) mathematics skills (2 scores).	1.7-2.4 2.5-3.4 3-8	1972	Standard - 212 min. Basic - 113 min. V & R - 72 min. L - 67 min. W - 80 min. M - 60 min.	G	MRC
IOWA TEST OF EDUCATIONAL DEVELOPMENT (Short Form)					
7 scores: Composite (which includes reading, language arts, mathematics), separate scores for reading, mathematics, language arts, social studies, science, use of sources.	9-12	1971	195 min.	J	Docutran
METROPOLITAN ACHIEVEMENT TEST					
Primer: Listening for sounds, Reading (word knowledge), Numbers.	K.7-1.4	1970	70 min.	F	IBM 805 IBM 1230
Primary I: 5 tests; Word Knowledge, Word Analysis, Reading, Mathematics Computations, Mathematics Concepts.	1.5-2.4		90 min.		NCS MRC
Primary II: 7 tests; Those tests listed in Primary I plus Spelling and Mathematics problem solving.	2.5-3.4		136 min.		
Elementary: 7 tests; listed in Primary II.					
Intermediate: 7 tests; Language test replaces Word Analysis.	5.0-6.9		270 min.		
Advanced: 9 tests; Social Studies and Science is added.	7.0-9.5		265 min.		

* Address given on Test Publisher Address List

ACHIEVEMENT TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
METROPOLITAN HIGH SCHOOL BATTERY 11 tests: Reading, Spelling, Language Arts, Language Study Skills, Social Studies Study Skills, Social Studies Vocabulary, Social Studies Information, Mathematics Computation and Concepts, Mathematical Analysis and Problem Solving, Scientific Concepts and Understandings, Science Information.	9-13	1962	315 min. (7 sittings)	F	MRC IBM 805 IBM 1230 Digitek
NATIONAL EDUCATIONAL DEVELOPMENT TEST 6 scores: English usage, Mathematics Usage, Social Studies Reading, Natural Science Reading, Word Usage, Total.	7-8 9-10	1970	180 min.	J	Docutran
SRA ASSESSMENT SURVEY Primary Edition Multi-Level ITED (Iowa Test of Educational Development) Broad knowledge, general skills, application, Reading, Language Arts, Mathematics. Multi-level and ITED also test Science, Social Studies, and use of Sources.	 1-4 4-9 9-12	1971	 2 hr., 30 min. 4 hr., 30 min. 3 hr., 15 min.	J	Docutran Hand-scoring
STS EDUCATIONAL DEVELOPMENT SERIES 14 reports and scores for grades 4-12: 3 interest areas (school plans, career plans, school interests); 3 ability scores (verbal, nonverbal, total); 7 achievement scores (reading, English, mathematics, basic skills total, science, social studies, everyday problems). Composite of ability and achievement scores. 9 reports and scores for grades 2-3 omitting school plans, career plans, science, social studies, and everyday problems.	2-3 4-6 6-9 9-12	1963-1971	150 min. 245 min. 235 min. 235 min.	H	NCS

* Address given on Test Publisher Address List.

ACHIEVEMENT TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
SEQUENTIAL TESTS OF EDUCATIONAL PROGRESS (STEP) SERIES II 7 areas: English Expression, Reading, Mechanics of Writing, Mathematics Computation, Basic Concepts, Science, and Social Studies.	4-14	1969	410 min.	D	Hand-scoring MRC IBM 805 IBM 1230 Digitek
STANFORD ACHIEVEMENT TEST Subtests in (grades 1.5-9.9) Reading, (grades 4.0-9.9) spelling and language, arithmetic, and (grades 5.5-9.9) science and social studies available as separates; partial batteries without science and social studies are available at Intermediate I and higher levels. High School Basic Battery 9 scores: English, Numerical Competence, Mathematics, Reading, Science, Social Studies, Spelling - 3 supplementary tests - Arts and Humanities, Business and Economics, Technical Comprehension.	1.5-2.4 2.5-3.9 4.0-5.4 5.5-6.9 7.0-9.9 9-12	1923-1968	127 min. 200 min. 269 min. 267 min. 255 min. 320 min.	F	MRC IBM 805 IBM 1230 Digitek Hand-scoring
TESTS OF ACADEMIC PROGRESS 7 scores: Social Studies, Composition, Science, Reading, Mathematics, Literature, Total - subtests available as separates.	9-12	1964-1966	330 min.	G	Digitek IBM 805 IBM 1230 MRC Hand-scoring
WIDE RANGE ACHIEVEMENT TEST (Revised Edition) 3 scores: Spelling, Arithmetic Reading. Individually Administered	Ages 5-0 to 11-11, 12-0, & over	1940-1965	20-30 min.	I	Hand-scoring

* Address given on Test Publisher Address List.

DIAGNOSTIC TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
CALIFORNIA PHONICS SURVEY Form 1, 9 error analysis scores: Long-short vowel confusion, other vowel confusion, consonant confusions with blends and diagraphs, consonant-vowel reversals, configuration, endings, negatives-oppositives-sight words, rigidity, total.	7-12 College	1956-1963	40-45 min.	C	Scoreze IBM 805
DIAGNOSIS: AN INSTRUCTIONAL AID Criterion-referenced: phonetic analysis, structural analysis, comprehension, vocabulary, use of sources.	1-4	1973	Not Reported	J	Hand-scoring
DIAGNOSTIC READING SCALES 10 or 11 scores: Word recognition, Instructional level (oral reading), Independent level (silent reading), Potential level (auditory comprehension) and 6 phonic scores: Consonant sounds, vowel sounds, consonant blends, common syllables, blends, and letter sounds.	1-8 9-12 (Reading Disability)	1963	20-30 min.	C	Hand
DIAGNOSTIC READING TESTS 3 levels: Survey Section, Word Attack, Vocabulary Rate of Reading.	K-4 4-8 7-13	1947-1963	90 min. 90 min. 165 min.	L	Hand IBM Answer Sheet
DURRELL LISTENING-READING SERIES 9 scores: Listening (vocabulary, sentences or paragraphs, total -- these scores are referred to as "potential reading grade equivalents"), Reading (vocabulary, sentences or paragraphs, total -- these scores are referred to as "actual reading grade equivalents"), potential minus actual reading grade equivalents (vocabulary, sentences or paragraphs, total -- these scores are referred to as "differential" scores).	1-2 3-6 7-9	1969-1970	140 min. (2 sessions, 2-7 days apart) 170 min. (2 sessions, 2-7 days apart) 160 min. (2 sessions, 2-7 days apart)	F	MRC Hand-Scoring Digitek IBM 805 IBM 1230

* Address given on Test Publisher Address List.

DIAGNOSTIC TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
GATES-McKILLOP READING DIAGNOSTIC TEST 28 Scores: Omissions, Additions, Repetitions, 8 mispronunciation scores, Oral reading, Word perception, Phrases, Blending word parts, Giving letter sounds, Naming letters, Recognizing visual forms of sounds, Auditory blending, Spelling, Oral Vocabulary, Syllabication, Auditory Discrimination.	2-6	1926-1962	30-60 min.	B	Not Reported
GATES-RUSSELL SPELLING DIAGNOSTIC TEST 9 scores: Spelling words orally, Work pronunciation, Giving letters for letter sounds, Spelling one syllable, Spelling two syllables, Word reversals, Spelling attack, Auditory discrimination, visual-auditory-kinesthetic.	2-6	1937-1940	Not Reported	B	Not Reported
GILMORE ORAL READING TEST (New Edition) 3 scores: Accuracy, Comprehension, Rate.	1-8	1951-1968	15-20 min.	F	Hand
GRAY ORAL READING TEST Reading passage booklet.	1-16	1963	Not Reported	A	Not Reported
McCULLOUGH WORD ANALYSIS TESTS (Experimental Edition) 10 scores: Phonetic (Initial blends and diagraphs, phonetic discrimination, matching letters to vowel sounds, sounding whole words, interpreting phonetic symbols, phonetic analysis total); Structural (dividing words into syllables, root words in affixed forms, structural-analysis total), Total.	4-6	1960-1963	70 min.	E	Not Reported

* Address given on Test Publisher Address List.

DIAGNOSTIC TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
PUPIL PROGRESS SERIES 4 levels in grades 1-8: Primary level, Vocabulary, Rate of Reading, Total Comprehension. Elementary and Advanced - Knowledge and Use of Sources, Rate of Reading, Reading Comprehension.	1-2 2-3 4-6 7-8	1965; 1968	45-50 min.	H	Primary-Hand-Scoring Elem. & Advanced - Hand-Scoring or STS score
SILENT READING DIAGNOSTIC TESTS 8 tests, 17 scores: Words in isolation, Words in context, Total right, Initial errors, Middle errors, Ending errors, Orientation errors, Total errors, Total omitted, visual-structural syllabication, Word synthesis, Total right, Beginning sounds, Ending sounds, Vowel and Consonant sounds.	2-6	1966-1970	90 min. in 2 or 3 sessions	M	Hand-Scoring
STANFORD DIAGNOSTIC ARITHMETIC TEST 2 levels: Concepts, computation, number facts, fractions.	2.5-4.5 4.5-8.5	1966-1968	170 min. in 6 sessions 210 min. in 7 sessions	F	Scoring Service
STANFORD DIAGNOSTIC READING TEST 2 levels: Comprehension, vocabulary, syllabication, auditory, discrimination, beginning and ending sounds, blending, sound discrimination, rate.	2.5-4.5 4.5-8.5	1966-1968	137 min. (4 sittings) 91 min. (3 sittings)	F	MRC Hand-Scoring

* Address given on Test Publisher Address List.

INTEREST TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
GORDON OCCUPATIONAL CHECK LIST Interest check list for non-college bound. It contains 240 statements of job activities.	8-12 Adult	1961-1967	20-25 min.	F	Hand-Scoring
GORDON PERSONAL INVENTORY 4 scores: Cautiousness, Original Thinking, Personal Relations, Vigor.	8-16 Adult		15-20 min.	F	Hand IBM
KUDER GENERAL INTEREST SURVEY Measures preferences in 11 areas: Outdoor, mechanical, scientific, computational, persuasive, artistic, literary, musical, social service, clerical, and verification.	6-12	1934-1970	45-60 min.	J	Self-Scoring Answer Pads or DocuTran
KUDER OCCUPATIONAL INTEREST SURVEY 77 occupational scores and 29 college-major scores are reported for men, and 57 occupational scores and 27 college-major scores for women.	11-16 Adult	1956-1970	30-40 min.	J	Machine Scored by Company
KUDER PERSONAL PREFERENCE SURVEY An individual's personal and social preferences as distinguished from occupational preference. The following types of social situations are reported: participation in group activities, familiar and stable situations, working with ideas, avoiding conflict and directions, and influencing others.	9 - Adult	1970	30-40 min.	J	Hand-Scoring
KUDER VOCATIONAL PREFERENCE RECORD Designed to measure interest in 10 areas: Mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, clerical, and outdoor.	9 - Adult	1970	30-40 min.	J	Self-Scoring Pin Punch Answer Pads

* Address given on Test Publisher Address List.

INTEREST TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
MINNESOTA VOCATIONAL INTEREST INVENTORY 30 scores: Males not planning to attend college. 21 occupational scales, 9 area scales	Age 15 & over	1965-1966	45-50 min.	I	Hand MKC NCS
OCCUPATIONAL INTEREST SURVEY (With Pictures) Individual Placement Series 9 scores: Scientific, social service, literary, agricultural, business, mechanical, musical, clerical, and artistic.	9-Adult	1959-1966	15-20 min.	N	Hand-Scoring
OHIO VOCATIONAL INTEREST SURVEY Reports students' interests along 24 scales by percentile ranks and stanines by grade and sex (280 items). Summary report presents numbers and percents of students choosing various options.	8-12	1969-1970	69-95 min.	F	MRC NCS
PICTURE INTEREST INVENTORY 9 scores: Interpersonal service, natural, mechanical, business, esthetic, scientific, verbal, time perspective, computational.	7-Adult	1958	30-40 min.	C	IBM 805 Hand or Machine
STS JUNIOR INVENTORY Measures 4 areas of interest: About Me and My School, About Myself, Getting Along With Others, Things in General.	4-8		35-40 min.	H	STS Machine Scoring or Hand
STS YOUTH INVENTORY ** Measures 5 areas of interest: My School, After High School, About Myself, Getting Along With Others, Things in General. ** Spanish Edition available.	7-12		30-35 min.	H	STS Machine Scoring or Hand

* Address given on Test Publisher Address List.

INTEREST TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
SCHOOL INTEREST INVENTORY Useful in identifying school dropouts (150 item scale).	7-12	1966	20-30 min.	G	MRC Hand
STRONG-CAMPBELL INTEREST INVENTORY	Age 17 or over	1976	No Limit. Approximately 30-40 min.	I	Hand or Machine
THURSTONE INTEREST SCHEDULE Yields profiles of interest in 10 fields: Physical Science, Biological Science, Computational, Business, Executive, Persuasive, Linguistic, Humanitarian, Artistical, Musical.	9-Adult	1947	10 min.	I	Hand
WHAT I LIKE TO DO: An Inventory of Children's Interest 8 scores: Art, Music Social Studies, Active Play, Quiet Play, Manual Arts, Science.	4-7		50-60 min.	J	Hand IBM 805

* Address given on Test Publisher Address List.

MENTAL ABILITY TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
ACADEMIC PROMISE TESTS 7 scores: Abstract reasoning, Numerical, Nonverbal Total, Language Usage, Verbal, Verbal Total, Total.	6-9		90 min.	I	MRC IBM 805 IBM 1230 OpScan
ANALYSIS OF LEARNING POTENTIAL (ALP) 5 batteries employ novel item types which are as free as possible for specific in-school instruction results: General capacity for school learning, Reading, Mathematics.	1-12	1970	49 min. 50 min. 68 min.	F	IBM 805 IBM 1230 Digitek MRC
CALIFORNIA SHORT-FORM TEST OF MENTAL MATURITY - 1963 Revision Logical Reasoning, Numerical Reasoning, Verbal Concepts, Memory, Language, Non-Language.	K-12	1963	34 min. 41 min. 42 min. 43 min. 41 min. 39 min.	C	Compu-Scan
CALIFORNIA TEST OF MENTAL MATURITY - 1963 Revision 12 tests divided into Language, Non-Language.	K-12	1963	48 min. 63 min. 83 min. 81 min.	C	IBM 1230
"FORM O" MENTAL ABILITY Measure quantitative, verbal, and pictorial reasoning ability. Total score is converted to mental age and deviation IQ.	2-8	1963	About 45 min.	H	Hand
HENMON-NELSON TESTS OF MENTAL ABILITY Standard group intelligence test.	3-6 6-9 9-12	Revised Edition	30 min.	G	IBM MRC

* Address given on Test Publisher Address List.

MENTAL ABILITY TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
KUHLMAN-ANDERSON INTELLIGENCE TESTS - 7th Edition 3 scores for Grades 7-12: Verbal, Quantitative, Total.	K-12	1960-1963	50-60 min.	I	Hand IBM 805 IBM 1230 OpScan
LORGE-THORNDIKE INTELLIGENCE TESTS Multi-Level Edition 2 Tests: Verbal and Non-Verbal.	3-13		35 min. 27 min.	G	MRC
OTIS-LENNON MENTAL ABILITY TEST Deviation IQ's, Percentile ranks, and stanines by age and grade.	K-12	1967	30 min. 45 min. 40 min.	F	MRC IBM 805 IBM 1230 Digitek NCS Hand
QUICK WORD TEST Measures Verbal intelligence, scores percentiles and stanines by grades for 4-12.	4-12 College Adult	1957-1967	Approximately 15 min.	F	IBM 805 IBM 1230 Digitek
PRE-PRIMARY MENTAL ABILITY Total score is converted to mental age and nonverbal IQ.	K-1		About 30 min.	H	Hand
SCHOOL AND COLLEGE ABILITY TESTS (SCAT) - Series II 3 scores: Verbal, Quantitative, Total.	4-14	1970	40 min.	D	IBM 805 IBM 1230 Digitek

* Address given on Test Publisher Address List.

MULTI-APTITUDE TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
DETROIT GENERAL APTITUDES EXAMINATION 3 Aptitude scores: Intelligence, Mechanical, Clerical. "Analysis of Results" chart is helpful in locating scores on emotional and social maladjustment.	7-12		85-90 min.	A	Hand
DIFFERENTIAL APTITUDE TESTS 7 booklets, 9 scores: Verbal Reasoning, Numerical Ability, Abstract Reasoning, Mechanical Reasoning, Clerical Speed and Accuracy, Language Usage, Spelling, Abstract Reasoning.	8-12 Adult	1973	190 min.	I	IBM 805 IBM 1230 OpScan MRC
FLANAGAN APTITUDE CLASSIFICATION TEST Measures aptitude for 19 skills essential to success in 37 broad occupational areas.	9-12 Adult		5-40 min. per test	J	Hand
MULTIPLE APTITUDE TEST 9 scores Factor 1. 2 scores: Word Meaning, Paragraph Meaning. Factor 2. 2 scores: Language Usage, Routine Clerical Facility. Factor 3. 2 scores: Arithmetic Reasoning, Arithmetic Computation. Factor 4. 3 scores: Applied Science and Mechanics, 2-dimensional Spatial Relations, 3-dimensional Spatial Relations.	7-13	1959	2 hours, 57 min.	C	Hand IBM 1230

* Address given on Test Publisher Address List.

PERSONALITY TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
THE ADJUSTMENT INVENTORY (Bell) Revised Student Form-6 scores: Home, Health, Masculinity, Emotionality, Submissiveness, Hostility Adult Form-6 scores: Home, Occupational, Health, Social, Emotional, Total.	9-16	1934-1963	20-30 min.	I	IBM
CALIFORNIA TEST OF PERSONALITY 1954 Revision 15 scores: Self-reliance, Sense of Personal Worth, Sense of Personal Freedom, Feeling of Belonging, Withdrawing Tendencies, Nervous Symptoms, Total Personal Adjustment, Social Standards, Social Skills, Antisocial Tendencies, Family Relations, School Relations or Occupational Relations, Community Relations, Total Social Adjustment, Total Adjustment.	K-3 4-8 7-10 9-16 Adult	1954	No limit 40-50 min. is usually sufficient.	C	Scoreze Machine
CHILDREN'S PERSONALITY QUESTIONNAIRE 14 scales: Leadership, Delinquency, Underachievements, etc.	2-7		50 min.	K	Hand
JR-SR HIGH SCHOOL PERSONALITY QUESTIONNAIRE Measures 14 independent factors: General intelligence, emotional maturity, anxiety level, & others.	7-12	1953-1969	Approximately 40 min.	A K	Hand Digitek
MINNESOTA COUNSELING INVENTORY 9 scores: Family Relationships, Social Relationships, Emotional Relationships, Conformity, Adjustment to Reality, Mood, Leadership, Validity, Question	9-12	1953-1957	50 min.	I	Hand Machine
MOONEY PROBLEM CHECK LIST 1950 Revision Check problems of some concern and write summary. Junior High: 7 scores, High School: 11 scores, College: 11 scores, Adult: 9 scores.	7-9 9-12 College Adult	1950	20-50 min.	I	Hand IBM

* Address given on Test Publisher Address List.

READING READINESS TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
CLYMER-BARRETT PRE-READING BATTERY 4 scores: Visual Discrimination, Auditory Discrimination, Visual-Motor, Total.	First Grade Entrants	1966-1969	90 min. in 3 sessions	E	Hand
GATES-MacGINITE READING TESTS: READINESS SKILLS TEST 9 scores: Listening Comprehension, Auditory Discrimination, Visual Discrimination, Following Directions, Letter Recognition, Visual-Motor Coordination, Auditory Blending, Total; Word Recognition.	K-1	1939-1969	120 min. in 4 sittings	B	Hand
HARRISON-STROUD READING READINESS PROFILES 5 group tests: Using Symbols, Making Visual Discriminations, Using the Context, Making Auditory Discriminations, Using Context and Auditory Clues. Optional individual test: Names of capitals and lower-case letters.	K-1	1949-1956	79 min. in 3 sittings	G	Hand
LEE-CLARK READING READINESS TEST 1962 Revision 4 scores: Letter Symbols, Concepts, Work Symbols, Total.	K-1	1931-1962	20 min.	C	Hand
METROPOLITAN READINESS TESTS 7 or 8 scores: Word Meaning, Listening, Matching, Alphabet Numbers, Copying, Total; Draw-A-Man (optional).	K-1	1933-1969	60 min. in 3 sessions	F	Hand Machine
MURPHY-DURRELL READING READINESS ANALYSIS 6 scores: Sound Recognition, Letter Names (capitals, lower case, total), Learning Rate, Total.	First Grade Entrants	1949-1965	60 min. in 2 sittings	F	Hand
STEINBACH TEST OF READING READINESS 4 tests: Letter Identification, Word Identification, Ability to Follow Directions, Ability to Relate Words and Pictures.	First Grade Entrants		Approximately 40 min.	H	Hand

* Address given on Test Publisher Address List.

T15

READING TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
ADULT BASIC READING INVENTORY 5 scores: Sight Words, Sound and Letter Discriminations, Word Meaning (reading), Word Meaning (Listening), Context Reading.	6-12 Adult		Approximately 60 min.	H	Hand
BURNETT READING SERIES Measures: Context clues, inferential reasoning, knowledge of grammatical structures, word meaning, word identification, rate and accuracy.	A: 1-12 B: 1-6.9		50 min. 50 min. 45 min. 45 min. 45 min.	H	STS Scoring Service (Grade 4 & above)
CALIFORNIA READING ACHIEVEMENT TESTS (1970 Edition) 5 levels: Each level comprises two sections (1) Vocabulary, which tests words in context, and (2) Comprehension, which tests ability to use reference materials and understand reading passages.	1.5-2 2-4 4-6 6-9 9-12	1970	46 min. 40 min. 45 min. 50 min. 50 min.	C	Compu-Scan (lower levels) IBM 1230 Digitek
COMPREHENSIVE TEST OF BASIC SKILLS - READING TEST 2 tests: Vocabulary-which measures ability to use words in context, and Comprehension - which measures understandings of stories, articles, poems, letters, and ads.	2.5-4 4-6 6-8 8-12		49 min. 52 min. 46 min. 41 min.	C	Hand (Level 1) Compu-Scan Digitek IBM 1230 Scoreze (Levels 2, 3, 4)
DAVIS READING TEST 2 scores: Level of Comprehension, Speed of Comprehension.	8-11 11-13		40 min.	I	Hand Machine
GATES-MacGINITE READING TESTS Primary A, B, C: Measures and comprehension. Primary CS: Speed and accuracy. Survey D or DM, E or EM, F or FM, 3 tests: Speed and accuracy, vocabulary, comprehension.	1-3 4-6 7-9 10-11	1965	A, B, C - 40 min. CS-7 min. D or DM-46 min. E or EM-45 min. F or FM-45 min.	I	Hand Machine MRC IBM 805

* Address given on Test Publisher Address List.

READING TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
IOWA SILENT READING TEST New Edition Measures: Vocabulary, reading comprehension, directed reading, reading efficiency, total.	6-9	1973	Level I: 1 hr., 31 min. Level 2: 1 hr., 26 min. Level 3: 56 min.	F	Hand IBM 805 IBM 1230 MRC Digitek OpScan
METROPOLITAN READING TESTS 1970 Revision The tests measure vocabulary and comprehension. Primary II, Elementary, Intermediate, and Advanced contain both work knowledge and reading subtests.	2.5-3.4 3.5-4.9 5.0-6.9 7.0-9.5	1970	48 min. 40 min. 40 min. 40 min.	F	MRC IBM 805 IBM 1230
MONROE READING APTITUDE TESTS 5 scores: Visual, Auditory, Motor Control, Oral Speed and Articulation, Language.	K-1		Group, 30-40 min. Individual items, 10-15 min.	G	Booklets
THE NELSON-DENNY READING TEST Revised Edition 4 scores: Vocabulary, Comprehension, Rate, Total.	9-16 College Adult		30 min.	G	MRC Hand IBM 805 IBM 1230
THE NELSON READING TEST Revised Edition Measures: Comprehension, general significance of a paragraph, important details, predict outcomes.	3-9		30 min.	G	MRC Answer-card Hand IBM 805 IBM 1230 Digitek
PRIMARY READING PROFILES ** Skills Measured: Aptitude for reading, Auditory association, Word recognition, Word attack, Reading comprehension. ** Also used as a Diagnostic Test	1.5-2.0 2.5-3.0		Approximately 83 min.	G	MRC Machine Scorable Booklets

* Address given on Test Publisher Address List.

READING TESTS
(Representative Samples)
(Continued)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
STANFORD HIGH SCHOOL READING TEST Measures ability to understand material read, judge what is implied, draw inferences.	9-12	1965-1966	40 min.	F	MRC IBM 805 IBM 1230
STANFORD READING TESTS Primary I: Word Reading, Paragraph Meaning, Vocabulary, Word Study Skills.	1.5-2.4	1964-1965	83 min.	F	Hand MRC IBM 805 IBM 1230 Harbor Card
Primary II: Word Meaning, Paragraph Meaning, Word Study Skills.	2.5-3.9		72 min.		
Intermediate I: Word Meaning, Paragraph Meaning.	4.0-5.4		40 min.		
Intermediate II: Word Meaning, Paragraph Meaning.	5.5-6.9		42 min.		
Advanced: Paragraph Meaning	7.0-9.9		30 min.		

* Address given on Test Publisher Address List.

T12

OTHER TESTS
(Representative Samples)

TEST DESCRIPTION	GRADE LEVEL	DATE	ADMINISTRATION TIME	* PUB.	SCORING
PUPIL RATING SCALE FOR LEARNING DISABILITIES Teacher-used screening instrument on: Auditory Comprehension, Spoken Language, Orientation, Motor Coordination, Personal-Social Behavior.	2-5		Untimed	I	Hand
REMOTE ASSOCIATES TEST Measure of ability to think creatively. Based on associative interpretation.	9-12 College Adult		40 min.	G	Hand
SURVEY OF STUDENT HABITS AND ATTITUDES 7 score diagnostic profile of 4 basic and 3 combined areas: Decay Avoidance, Work Methods, Study Habits, Sub-Total, Teacher Approval, Education Acceptance, Study Attitudes Sub-Total, Total Study Orientation Score.	7-12 College		Untimed About 20-25 min.	I	Hand IBM 805 IBM 1230
TORRANCE TESTS OF CREATIVE THINKING Measures creativity by assessing 4 mental characteristics: Fluency, Flexibility, Originality, Elaboration.	K-3 (Indiv.) 4- College (Group)		Verbal Testing - 1 hour Figural Testing- 45 min.	E	Hand
WATSON-GLASER CRITICAL THINKING APPRAISAL Measures 5 aspects of critical thinking: Drawing sound inferences, recognizing assumptions, reasoning logically by assumption and interpretation, discriminating between arguments.	9-12 College Adult		Untimed About 50 min.	K F	Hand

* Address given on Test Publisher Address List.

SECTION 4

Information Sources

(1) Professional Associations -- In addition to joining the National Council on Measurement in Education, membership in other professional associations according to particular needs should be considered. Several of these follow:

National Council on Measurement in Education (NCME)
Office of Evaluation Service
Michigan State University
East Lansing, Michigan 48823

American Educational Research Association
1126 Sixteenth Street, N.W.
Washington, D.C. 20036

American Personnel and Guidance Association (APGA)
1607 New Hampshire Avenue, N.W.
Washington, D.C. 20009

American Psychological Association
1200 Seventeenth Street, N.W.
Washington, D.C. 20036

(2) Newsletters, Special Publications, and Professional Journals -- Each of the associations listed above sponsors publications which will be useful. Among them are MEASUREMENT NEWS (NCME); AMERICAN EDUCATIONAL RESEARCH JOURNAL (AERA); REVIEW OF EDUCATIONAL RESEARCH (AERA); EDUCATIONAL RESEARCHER (AERA); MEASUREMENT AND EVALUATION IN GUIDANCE (AMEG). Another important journal is EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina 27703.

(3) Useful References -- Below are listed some references which range from the popular and practical to the technical and theoretical. As noted above, they are intended to reflect the spectrum of demands upon the test director.

Ahman, J. Stanley and Marvin D. Glock. Evaluating Pupil Growth. 5th Edition. Boston: Allyn and Bacon, Inc. 1975.

American Psychological Association. Standards for Educational and Psychological Tests and Manuals. Washington, D.C. 1974.

Anastasi, Anne (Ed.). Testing Problems in Perspective. Washington, D.C.: American Council on Education. 1966.

Bauernfeind, R. H. Building A School Testing Program. Boston: Houghton Mifflin Company. 1969.

Bloom, B. B., Hastings, J. T., and Madaus, G. F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill. 1971.

Ebel, Robert L.; Noll, Victor H.; and Bauer, Rober M. (Ed's). Encyclopedia of Educational Research, Fourth Edition. London: The MacMillan Company. 1969.

Gulliksen, Harold. Theory of Mental Tests. New York: John Wiley & Sons, Inc. 1965.

House, Ernest R. School Evaluation: The Politics and Process. Berkeley, Calif.: McCutchan. 1973.

Lyman, Howard B. Test Scores and What They Mean, Second Edition. Englewood Cliffs, N.J.: Prentice-Hall, Inc. 1971.

Russell Sage Foundation. Guidelines for the Collection, Maintenance, and Dissemination of Pupil Records. New York: Russell Sage Foundation. 1970.

(4) External Testing Programs -- Although the test director may have no direct responsibility for testing programs sponsored by outside agencies, parents and teachers probably will ask him for information from time to time. If the office files do not contain up-to-date information on these several agencies and the programs they sponsor, the new test director will want to write to the following addresses:

American College Testing Program
P.O. Box 168
Iowa City, Iowa 52240

Educational Records Bureau
160 Speen Street
Framington, Massachusetts 01701

College Entrance Examination Board
Box 592
Princeton, New Jersey 08540

General Educational Development Testing Services
American Council on Education
1785 Massachusetts Avenue
Washington, D.C. 20036

National Assessment of Educational Progress
300 Lincoln Tower
1860 Lincoln Street
Denver, Colorado 80203

National Merit Scholarship Corporation
990 Grove Street
Evanston, Illinois 60201

TEST INSTRUMENTS LIBRARY

Located in the Planning, Evaluation, and Research Section of the Nebraska State Department of Education is the Test Instruments Library. The Library is an accumulation of approximately 1,000 specimen tests covering most subject areas and all grade levels (K-12). The Library also contains catalogs and miscellaneous material relative to those

specimen sets.

The Library contains many specimen tests in the various subject areas which are listed below:

Achievement Batteries
Affective
Character & Personality
English
Fine Arts
Foreign Language
Intelligence
Mathematics
Miscellaneous
Multi-Aptitude Batteries
Reading
Science
Sensory-Motor
Social Studies
Special Education
Speech & Hearing
Vocations

Housed in the Library are catalogs and ordering information.

Mental Measurements Yearbook -- gives details concerning measurements materials, researchers evaluate the measures.

CSE-Test Evaluation Booklets -- according to certain criteria, CSE has graded tests, Good - Fair - Poor.

ETS-Test Collection Bulletin -- ETS reports new tests, what they measure, and where to obtain them.

Microfiche is easily accessible and some tests are on microfiche. These tests are all public domain and there is no charge for using them.

HOW TO USE THE LIBRARY

If a person wants to review a specific specimen set(s) or general subject area(s), they should write a detailed letter of request and send it to: 301 Centennial Mall South, P.O. Box 94987, Lincoln, Nebraska 68509; or call 471-2486. Then the request will be thoroughly researched.

After the request is filled and notification is made, users are asked to come to the Department to use the Library. (Note: There is only one copy of each specimen set.) If this is impossible, as in most cases, then the material will be sent by mail. If sent, the policy is

to allow the person to keep the material a maximum of one month, unless there appears to be a great demand. The actual purchase of tests from a testing company for a testing program is up to the individual schools or districts.

SECTION 5

Computing Guides

The procedures of assessment are the procedures of the science of measurement, of estimating quantities and interpreting variation in numbers. The computing guides cover a wide range of computations which could be made, or are required, in the quantitative analysis of the measurement results.

Computing Guide 1 . . . . . Determining Percentile Equivalents

Computing Guide 2 . . . . . How to Compute Local Norms

Computing Guide 3 . . . . . Calculation of Standard Scores

Computing Guide 4 . . . . . Relations Between Standard Scores and Percentile Scores When Raw Scores Are Normally Distributed

Computing Guide 5 . . . . . Method for Estimating Expected Scores

Computing Guide 6 . . . . . Use of Spearman-Brown Formula

Computing Guide 7 . . . . . Internal Consistency Formulas

Computing Guide 8 . . . . . Stanine Computation for Sample Mathematics Scores

Computing Guide 9 . . . . . Reliability of Differences Between Part Scores

Computing Guide 10 . . . . . Standard Error of Measurement and Confidence Bands

Computing Guide 1 . . . . . Determining Percentile Equivalents[2]

1. Begin with the raw scores (these are scores of 75 ninth-grade boys on Bennett Form AA).

37	43	27	44	27	27	26	31	35	42	50
35	43	36	26	50	47	36	26	32	32	38
36	21	24	40	39	35	38	36	38	21	17
26	35	22	18	50	30	38	50	16	45	8
34	26	34	28	41	27	39	41	30	23	33
22	31	36	40	54	24	22	8	33	42	41
41	31	34	36	32	20	22	34	41		

2. Identify the highest score and the lowest score. If there is a wide range, choose a class interval of 1, 2, 5, 10, 20, etc., and divide the range into classes of equal width. Fifteen or more classes are desirable.

Highest score = 54; lowest score = 8; range = 46. Class interval of 5 will be used. (A smaller interval, such as 2, would be preferable but would be inconvenient in this computing guide.)

3. Tally the number of cases with each score.

4. Write the number of tallies in the Frequency (f) column. Add this column to get N, the number of cases.

5. Begin at the bottom of the column and add frequencies one at a time to determine the cumulative frequency, the number of cases below each division point.

6. Divide the cumulative frequencies by N to determine cumulative percentages.

Scores	Tallies	Frequency (f)	Cumulative Frequency	Cumulative Percent
			75	100
50-54	𝍸	5	70	93
45-49	//	2	68	90
40-44	𝍸 𝍸 //	12	56	75
35-39	𝍸 𝍸 𝍸 //	17	39	52
30-34	𝍸 𝍸 ////	24	25	33
25-29	𝍸 𝍸	10	15	20[b]
20-24	𝍸 𝍸	10	5[a]	7
15-19	///	3	2	3
10-14		0	2	3
5- 9	//	2	0	0
		75 N		

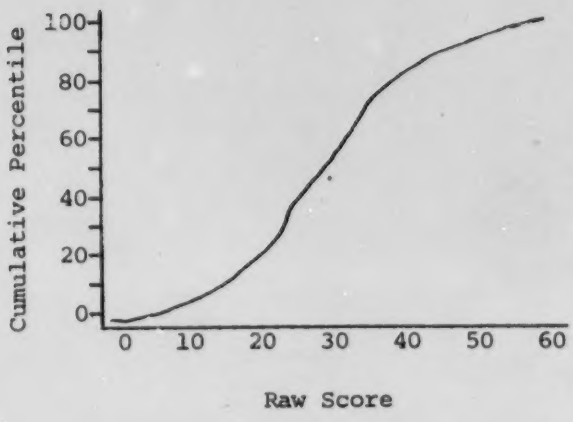
7. Plot cumulative percentage against score. (In practice, a large sheet of graph paper should be used.)

[a] 5 cases fall below 19.5; 15 below 24.5; etc.

[b] 20 percent of the cases fall below 24.5; 20 is the cumulative percentage corresponding to a raw score of 24.5.

[2] Quoted from Essentials of Psychological Testing, Cronbach, pg. 74-75.

8. Draw the smooth curve which best fits the points plotted.



9. Determine the percentile equivalent of a score by reading from the curve. (The lines on the chart show how one finds that the percentile equivalent of a raw score of 40 is 74.)

Percentile Equivalents

Raw	%ile	Raw	%ile	Raw	%ile
12	2	24	17	36	60
13	2	25	20	37	64
14	3	26	22	38	67
15	3	27	26	39	71
16	4	28	29	40	74
17	5	29	33	41	77

Computing Guide 2 . . . . . How To Compute Local Norms

The first step in the construction of local norms is to determine the specific groups upon which the norms are to be based. It is assumed that, when a large testing program is set up, the schools and students involved in the program have some common characteristics. Therefore, computing norms for all the participants in the program seems to be an immediately useful step. In addition, school systems, individual schools, and even departments within schools in the program may want to compute norms for their own groups of students.

Once the groups upon which the norms are to be based have been determined, the next step is to obtain the test scores of all students in the group. In constructing norms, it is desirable to have the scores of at least 100 students. Norms can be computed for smaller groups of students but the larger the group, the more reliable the local norms will be.

After the scores of all students in the group are collected, there are two main things to be done in constructing the norms: a frequency distribution of the scores must be made, and the percentile rank of each score or score interval must be computed. We shall discuss each of these separately.

Making A Frequency Distribution

Look through the scores to find the highest and the lowest score. Set up intervals of scores running from the highest down to the lowest score.

These intervals should be equal in size and there should be about 25 intervals. Set up as many intervals as there are possible scores on the test allowing one score per interval or set up as few as two intervals. However, if too many intervals are established, the norms table will be long and difficult to use.

An example may help to clarify the procedure. Suppose that we have the test scores of 300 students which we wish to use in developing local norms. The highest score earned on the test is 299 and the lowest score is 200. As shown in the first column of the sample work sheet, we would set up and list 25 score intervals, each interval containing four scores. Care should be taken that no intervals overlap; e.g., if the two intervals 240-243 and 243-246 were listed, we would not know where to put a score of 243.

After the score intervals have been determined and listed, a tally should be made of the number of students earning scores falling in each score interval. On the sample work sheet, this has been done for the fictitious group of 300 students mentioned above, and the number of students whose scores fall into each interval is listed in the column headed "Frequency." This third column, then, is the frequency distribution of scores.

Computing Percentile Ranks

The next job is to compute the percentile rank of each score interval. Percentile ranks are most easily computed from cumulative frequencies.

These are obtained by adding frequencies in the score distribution from the bottom up so that the number opposite each score interval equals the sum of the frequency for that interval and the frequencies of all of the intervals below it. The number opposite the top interval should equal the total number of students in the norms group. The cumulative frequencies have been computed for our imaginary group and posted on the sample work sheet (column 4).

The percentile rank for any score interval is determined by:

(a) Finding one-half the frequency for that score interval.
(b) Adding the result of (a) to the cumulative frequency for the score interval just below the interval in question.
(c) Dividing the result of (b) by the total number of students in the norms group (taking the answer to the nearest hundredth).
(d) Multiplying the answer from (c) by 100.

SAMPLE WORK SHEET

Score Interval	Tally	Frequency	Cumulative Frequency	Percentile Rank
269-299	/	1	300	99*
292-295		0	299	99*
288-291	/	1	299	99*
284-287	//	2	298	99
280-283	///	3	196	98
276-279	𝍸	5	293	97
272-275	𝍸 ///	8	288	95
268-271	𝍸 𝍸 ///	13	280	91
264-267	𝍸 𝍸 𝍸 /	16	267	86
260-263	𝍸 𝍸 𝍸 ////	19	251	81
256-259	𝍸 𝍸 𝍸 𝍸 /	21	232	74
252-255	𝍸 𝍸 𝍸 𝍸 //	22	211	67
248-251	𝍸 𝍸 𝍸 𝍸 ///	23	189	59
244-247	𝍸 𝍸 𝍸 𝍸 ///	23	166	52
240-243	𝍸 𝍸 𝍸 𝍸 //	22	143	44
236-239	𝍸 𝍸 𝍸 𝍸 /	21	121	37
232-235	𝍸 𝍸 𝍸 ////	19	100	30
228-231	𝍸 𝍸 𝍸 ///	18	81	24
224-227	𝍸 𝍸 𝍸 /	16	63	18
220-223	𝍸 𝍸 ////	14	47	13
216-219	𝍸 𝍸 //	12	33	9
212-215	𝍸 𝍸	10	21	5
208-211	𝍸 //	7	11	3
204-207	///	3	4	1
200-203	/	1	1	1*

Total number of students = 300

Upper quartile = 75th percentile = 258.2

Median = 59th percentile = 244.7

Lower quartile = 25th percentile = 230.2

Mean = 244.2

*Percentile ranks higher than 99 or lower than 1 may be recorded as 99 and 1, respectively.

This has been done for each score interval on the sample work sheet for the group of 300 students. Thus, for the 212-215 interval, the follow- computations were performed to obtain the percentile rank of scores falling in that interval:

(a) ½ x 10 = 5
(b) 5 + 11 = 16
(c) 16 ÷ 300 = .05
(d) .05 x 100 = 5

For the score interval, 256-259, a percentile rank of 74 was computed as follows:

(a) ½ x 21 = 10.5
(b) 10.5 + 211 = 221.5
(c) 221.5 ÷ 300 = .74
(d) .74 x 100 = 74

The percentile ranks listed on the right-hand side of the worksheet, then, are the norms for our fictitious group of 300 students. The procedure outlines above can be used to obtain program norms or local norms based on any desired group of students.

Computing Some Summary Statistics: Quartiles, Median, and Mean

The scores corresponding to the upper quartile (75th percentile), median (50th percentile), and lower quartile (25th percentile) divide the group into four quarters. To compute the upper quartile, these steps can be carried out:

(a) Take 75% of the total number of the group.
(b) Look in the cumulative frequency column for the nearest number smaller than the result of (a).
(c) Subtract the number obtained in (b) from the result of (a).
(d) Divide the result of (c) by the frequency of the interval next above the interval in which (b) was found.
(e) Multiply the result of (d) by the number of scores in a score interval.

(f) Add the result of (e) to the upper limit of the score interval in which (b) was found.*

To obtain the median and lower quartile, perform the same operations EXCEPT for (a). To compute the median, substitute 50% for 75% in (a). To compute the lower quartile, substitute 25% for 75% in (a).

Here is how the upper quartile, median, and lower quartile were obtained for the sample problem.

Upper Quartile

(a) 75% of 300 = 225
(b) 211 is the number in the cumulative frequency column which is the nearest number smaller than 225.
(c) 225 - 211 = 14
(d) 14 ÷ 21 = 2/3
(e) 2/3 x 4 = 8.3 or 2.7
(f) 2.7 + 255.5 = 258.2

Median

(a) 59% of 300 = 150
(b) 143 is the number
(c) 150 - 143 = 7
(d) 7 ÷ 23 = 7/23
(e) 7/23 x 4 = 28/23 or 1.2
(f) 1.2 + 243.5 = 244.7

Lower Quartile

(a) 25% of 300 = 75
(b) 63 is the number
(c) 75 - 63 = 12
(d) 12 ÷ 18 = 12/18 or 2/3
(e) 2/3 x 4 = 8/3 or 2.7
(f) 2.7 + 227.5 = 230.2

All teachers are familiar with the average or arithmetic mean, which is usually computed by summing all the scores for a group and then dividing by the number of individuals in the group. A mean can be computed directly from a frequency distribution by this method:

(a) Multiply the midpoint of each score interval by the corresponding frequency.

---

*The upper limit of a score interval is a value halfway between the last number written for that interval and the first number written for the next interval. For example, the upper limit for the interval 244-247 is 247.5, the upper limit for the interval 248-251 is 251.5, and the upper limit for the interval 252-255 is 255.5.

(b) Sum the products determined in (a).
(c) Divide the result of (b) by the total number of students in the group.

To obtain the mean for the sample distribution, these operations were carried out:

(a)
297.5 x 1 = 297.5
293.5 x 0 = 0
289.5 x 1 = 289.5
285.5 x 2 = 571.0
281.5 x 3 = 844.5
277.5 x 5 = 1387.5
273.5 x 8 = 2188.0
269.5 x 13 = 3503.5
265.5 x 16 = 4248.0
261.5 x 19 = 4968.5
257.5 x 21 = 5407.5
253.5 x 22 = 5577.0
249.5 x 23 = 5738.5
245.5 x 23 = 5646.5
241.5 x 22 = 5313.0
237.5 x 21 = 4987.5
233.5 x 19 = 4436.5
229.5 x 18 = 4131.0
225.5 x 16 = 3608.0
221.5 x 14 = 3101.0
217.5 x 12 = 2610.0
213.5 x 10 = 2135.0
209.5 x 7 = 1466.5
205.5 x 3 = 616.5
201.5 x 1 = 201.5
(b) Total = 73274.0
(c) 73274 ÷ 300 = 244.2

Of course, if a calculating machine is available, (a) and (b) can be performed in one operation with the "cumulative" multiplier.

Computing Guide 3 . . . . . Calculation of Standard Scores[3]

1. Begin with the raw scores to be converted, and find the mean and s.d. as in Computing Guide 2.

   Sample data:

   $M = 33.2$, s.d. $= 9.80$

2. To obtain z-scores, express each raw score as a deviation from the mean. Divide by the s.d.

   $$z \text{ score} = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

   For raw score 50:

   $$z = \frac{50 - 33.2}{9.80} = \frac{16.8}{9.80} = 1.7$$

   For raw score 25:

   $$z = \frac{25 - 33.2}{9.80} = \frac{-8.2}{9.80} = -.8$$

3. To obtain T scores, multiply the z score by 10 and add to 50.

   $$T \text{ score} = 50 + \frac{10\ (\text{raw score} - \text{mean})}{\text{standard deviation}}$$

   For raw score 50, $z = 1.7$

   $T = 50 + 10\ (1.7) = 67$

   For raw score 25, $z = -.8$

   $T = 50 + 10\ (-.8) = 42$

[3]Quoted from Essentials of Psychological Testing, Cronbach, pg. 81.

Computing Guide 4 . . . . . Relations Between Standard Scores and Percentile Scores, When Raw Scores are Normally Distributed[4]

Distance from Mean in s.d. (z score)	T Score	Percentile Score	Percent of Cases in "Tail" of Curve	Percentile Score	T Score	Distance from Mean in s.d. (z score)
3.0	80	99.9	0.1	0.1	20	-3.0
2.9	79	99.8	0.2	0.2	21	-2.9
2.8	78	99.7	0.3	0.3	22	-2.8
2.7	77	99.6	0.4	0.4	23	-2.7
2.6	76	99.5	0.5	0.5	24	-2.6
2.5	75	99.4	0.6	0.6	25	-2.5
2.4	74	99.2	0.8	0.8	26	-2.4
2.3	73	99	1	1	27	-2.3
2.2	72	99	1	1	28	-2.2
2.1	71	98	2	2	29	-2.1
2.0	70	98	2	2	30	-2.0
1.9	69	97	3	3	31	-1.9
1.8	68	96	4	4	32	-1.8
1.7	67	96	4	4	33	-1.7
1.6	66	95	5	5	34	-1.6
1.5	65	93	7	7	35	-1.5
1.4	64	92	8	8	36	-1.4
1.3	63	90	10	10	37	-1.3
1.2	62	88	12	12	38	-1.2
1.1	61	86	14	14	39	-1.1
1.0	60	84	16	16	40	-1.0
0.9	59	82	18	18	41	-0.9
0.8	58	79	21	21	42	-0.8
0.7	57	76	24	24	43	-0.7
0.6	56	73	27	27	44	-0.6
0.5	55	69	31	31	45	-0.5
0.4	54	66	34	34	46	-0.4
0.3	53	62	38	38	47	-0.3
0.2	52	58	42	42	48	-0.2
0.1	51	54	46	46	49	-0.1
0.0	50	50	50	50	50	0.0

[4]Quoted from Essentials of Psychological Testing, Cronbach, pg. 85.

Computing Guide 5 . . . . . Methods for Estimating Expected Scores

The simplest, and easiest to interpret, score is the percentage of correct responses to the total number possible. This same formula can be used for overall scores or for subdivision scores for any breakdown desired. It is mathematically equivalent to the miles per gallon ratio; the difference in the resulting quotients is the difference in the units of measurement between the divisor and the dividend. Because the correct response and the total possible are counted in the same units, items, and because the number correct can never exceed the total possible, the ratio takes on the special properties of percentages.

Individual scores should be averaged for the group. But mean test scores by themselves are not interpretable even when broken down by population. They must be compared against some independently motivated criteria that represents an expected level of performance. This additional criteria signifies the context of interpretation and is wholly necessary to the interpretation of the test scores. It should be defined by a group decision making process that is reasonably representative of equally competent educators.

One method that can be used to obtain consensus on a standard of interpretation goes as follows. Each member of the working group would review each item and estimate how many of the respondents will answer the item correctly, or assign a percentage to the item indicating what that member of the group feels is the minimum acceptable number of respondents answering the item correctly, or both. In the first case, the object is to predict, as accurately as is subjectively possible, the number of respondents who will answer the item correctly out of the

total number of respondents, given knowledge of the item and the capabilities of the respondents. In the second case, the object is to mark the point below which the percentage of respondents answering the item incorrectly would be a matter of concern. The range above this cut-off point indicates those percentages which are adequate or satisfactory in the judgment of the group member.

The different estimates by the members of the group for each item can be averaged and the mean computed for the predicted "true" percentages, and the minimum acceptable percentages. The expected overall scores and expected subdivision score, for both kinds of estimates of expected performance, is the sum of the mean percentages for each item figured into the particular score.

For example, if the mean percentages of minimum acceptable performance on three items were .84, .66, .80, the expected score on a subdivision consisting of these three items only would be 2.3 on a number scale from 0 to 3. The only statistic that should be compared against the expected scores are the group means.

The tables of results prepared for interpretation would present the expected true score, the expected minimum acceptable score, and the actual group scores of the populations to whom the test was administered. The tables could include this information for overall scores, each subdivision score, and for each item. The group that formulates the independently motivated criteria should examine the tables and report out their studied reactions to actual results. The summary of findings would detail any noteworthy discrepancies between expected scores and the actual group scores.

Determining what actions or decisions should be undertaken or made on the basis of actual results is primarily a management problem. Even though the data can be validly obtained, it does not in itself recommend the best decision or response. Obviously where actual results fall well below the minimum acceptable cut-off points, something should be decided. The first questions would be, are there any plausible explanations for the discrepancy? Which explanations are the most likely? The second questions would be, what things could be done to reduce the discrepancy? Which are the best? What are the consequences of making no changes? These are matters of judgment, yet easily within the responsibility of the educational profession.

Computing Guide 6 . . . . . Use of Spearmen-Brown Formula[5]

1. Suppose that a test has a known reliability. The Spearman-Brown Formula, given at right, estimates the reliability of the score from a similar test n times as long.

$$r_n = \frac{nr}{1 + (n - 1)r}$$

where
r is the original reliability; r_n is the reliability of the test n times as long

2. To predict the reliability of a test twice as long as the original test, substitute in the formula n = 2.

If r = .40, $r_2 = \frac{2(.40)}{1 + (1).40}$

$$= \frac{.80}{1.40} = .57$$

3. Suppose the original test is to be reduced to only half its original length. The reliability of the short test is estimated using n = ½

$$r_{½} = \frac{½(.40)}{1 + (½ - 1)(.40)}$$

$$= \frac{.20}{1 - ½(.40)} = \frac{.20}{.80} = .25$$

---

[5]Quoted from Essentials of Psychological Testing, Cronbach, pg. 131.

Computing Guide 7 . . . . . Internal Consistency Formulas[6]

K-R + 20:

$$r_{xx} = \frac{n}{n-1}\left[1 - \frac{\Sigma pq}{Sx^2}\right]$$

K-R 21:
(Assume items of equal difficulty)

$$r_{xx} = \frac{n}{n-1}\left[1 - \frac{\bar{X}(n-\bar{X})}{nSx^2}\right]$$

Coefficient Alpha

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\Sigma Si^2}{Sx^2}\right]$$

Where:

n = number of items in test

p = proportion of people who answered item correctly. (If, for example, on Item 1, 6 of 30 people answered the item correctly, p for this item would be 6/30=.20)

q = proportion of people who answered item incorrectly (q = 1 - p)

pq = variance of a single item scored dichotomously (right or wrong)

Σ = summation sign indicating that pq is summed over all items

S_x^2 = variance of the total test

$\bar{X}$ = mean of the total test

S_i^2 = variance of single item

---

[6] Quoted from Measurement and Evaluation in Education and Psychology, Mehrens and Lehmann, pg. 113.

Computing Guide 8 . . . . Stanine Computation for Sample Mathematics Test Scores

Mathematics Test Score	Stanine
95	9
93	
93	
92	8
88	
86	
83	
83	7
82	
80	
77	
76	
75	
72	
71	6
70	
68	
66	
64	
64	
64	
63	
61	
59	
58	
55	5
Mdn	
55	
54	
53	
50	
49	
49	
47	
45	4
45	
42	
42	
40	
39	
36	
35	
34	3
34	
31	
31	
29	
24	
21	2
19	
17	1

Steps:

1. Find the median (see point marked Mdn.).
2. Count off 5 scores (10% of 50) on each side of the median. These values all have a stanine of 5.
3. Since 17% of 50 is 8.5, we can count off either 8 or 9 more scores on each side and assign values of 4 and 6. (We chose 9.)
4. Count off 6 more scores (12% of 50) on each side and assign values 3 and 7.
5. Since 7% of 50 is 3.5, we can count off either 3 or 4 more scores. (We should choose 4 to avoid splitting the two 93's into different stanines) on each side and assign values of 2 and 8.)
6. The two end values receive stanines of 1 and 9, respectively.
7. Due to rounding, this distribution does not give the exact proportions found in a stanine distribution, but it is close enough.

[7] Quoted from Measurement of Evaluation in Education and Psychology, Mehrens and Lehman, pg. 154.

Computing Guide 9 . . . . . Reliability of Differences Between Part Scores[8]

The reliability of a difference between two standard scores A and B is calculated by this formula:

$$r_{(A-B)\ (A-B)} = \frac{r_{AA} + r_{BB} - 2r_{AB}}{2 - 2r_{ab}}$$

When tests have low reliabilities or a high degree of overlap, the difference is highly unreliable.

---

[8]Quoted from Essentials of Psychological Testing, Cronbach, pg. 287.

Computing Guide 10 . . . . . The Standard Error of Measurement and Confidence Bands

The standard error of measurement, S_e, is given by the following formula:

$$S_e = S_x \sqrt{1 - r_{xx}}$$

Where r_{xx} is the reliability coefficient of measure x and,

S_x is the standard deviation of obtained scores on measure x.

S_x can be taken from the norm group administration, but it could also be calculated from locally collected data.

Confidence Bands

The interpretation of the numbers is governed by length of scale, from low to high, the type of score employed, and the intuition of something having significantly more or less quantity. Graphic displays heighten this intuition. The standard error of measurement restricts it. The interpreter must consistently recognize the effects of chance when reading results as point scores. The actually obtained point scores are only one of many very probable scores should the measure be administered again, even under nearly identical circumstance. Due to chance alone, an individual will score several points higher or lower on any given retest. After the display of the scores, the first piece of numerical information of which the reader should be made keenly aware is the computed standard error of measurement for the test. The range created by adding and subtracting the standard error from the actual point score

also marks the range of scores within which the individual's other possible scores would fall 64% of the time. The range of scores defined by 80% confidence, four out of five chances, is computed by subtracting and adding 1.3 times the standard error from the actual point score. The clearest display of the confidence band is done graphically by charting the upper and lower scores of a confidence band of some higher level, e.g., 80%, 75%, etc.

SECTION 6

Glossary

ACADEMIC APTITUDE - The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called scholastic aptitude, school learning ability, academic potential.)

ACHIEVEMENT - An indication of accomplishment of a skill or an activity.

ACHIEVEMENT TEST - A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills - usually as a result of planned instruction or training.

AGE NORMS - Scores representing the typical or average performance for persons in successive age groups. Such norms are generally used in the interpretation of mental ability test scores.

APPLIED PERFORMANCE TESTING - A field of criterion referenced measurement that employs some form of performance sampling or simultation as to stimulus for the measurement occasion. Particularly suited for the assessment of competency.

APTITUDE - A combination of abilities and other characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly called mental ability or intelligence tests); those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning, and are used to predict future performance--usually in a specific field, such as foreign language, shorthand, or nursing.

APTITUDE TEST - This type of test includes those of general academic ability, and those of special abilities such as verbal, numerical, mechanical, musical, etc., which measure both ability and previous learning and are used to predict future performance.

ASSESSMENT - A scientific collection of information that will answer questions and support informational reasoning about some particular object, state of affairs, past, present, or future, or of individuals, their accomplishments, present capabilities and future circumstance. When collected periodically through a standard procedure, it serves as accounting (or baseline description) function. When collected in order to provide information known to have bearing on decisions about the best or probable future circumstance for individuals or groups of individuals, by anyone, the information serves diagnostic function. The classification of individuals from diagnostic information should always be connected directly with courses of action. When the collection of information is singularly focused upon one individual, or group, or upon a program, procedure or policy decision, an evaluative function is being performed.

BATTERY - A group of several tests whose results are of value individually, in combination, and/or totally.

BIVARIATE CHART (Bivariate distribution) - A diagram in which a tally mark is made to show the scores of one individual on two variables. (The intersection of lines is determined by the horizontal and vertical scales from cells in which the tallies are placed.) Such a plot provides frequencies for the two distributions, and portrays the relation between the two variables as a basis for computation of the product-moment correlation coefficient.

CEILING - The highest level of ability measured by a particular test.

CLASS LIST - An alphabetical list report of pupils' scores on the various subtests taken. Pupils' scores can be grouped by class, or by the total grade, whichever is preferred.

COMPETENCY - A behavioral capability that may be attributed to individuals should they show, by past accomplishment and present performance, that they are able to perform the tasks and behaviors associated with the capability. May be taken both as object of instruction and as an object of measurement. Sets of competencies should be methodically defined operationally and will tend to exhibit hierarchial, sequential and class inclusion relationships.

COMPOSITE SCORE - A score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the composite. Most commonly, such scores are used for predictive purposes and the several weights are derived through multiple regression procedures.

CORRECTION FOR GUESSING - A technique devised to adjust the number of right answers for a test by using some portion of the wrong answers. The assumption is made that if an examinee guesses on an objective test, the number of right and wrong answers resulting will be proportional to the number of alternate responses in each item. Common correction formulas are:

$$R - \frac{W}{2},\ R - W,\ R - \frac{W}{2}$$

CORRELATION COEFFICIENT - An expression of the degree of relationship between two sets of measures for the same group of individuals. Correlation coefficients range from .00, denoting a complete absence of relationship to +1.00 and to -1.00, indicating a perfect positive or negative correspondence, respectively. The letter r is commonly used to refer to the correlation coefficient.

CRITERION - A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to measure, to predict, or to correlate with.

CRITERION-REFERENCED (Measurement, test) - The construction and scoring of psychological tests which yield scores that are interpreted on scales which are derived directly from properties of subjects' performance or from the logical properties of the stimulus to which the individual must respond. Notice that a scale does not have to contain special point scores (minimum competency level, mastery standards) to be criterion-referenced, although most criterion-referenced scales are designed to be used in conjunction with severing point scores. Criterion-referenced measures vary principally in the manner in which they conceive their objects of measurement which vary in amount of technical precision to which they lend themselves or in the manner in which the learning outcome is stated.

CRITERION-REFERENCED TEST - A test designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of what the student knows or can do, rather than in their relation to the scores made by some external reference group.

DESIGN, DESIGN STATEMENT (Assessment) - The plan, in standing document form, of an assessment program, identifying who will be assessed, when, and by what instrument (the administrations chart), how, and to what specific use. It may also contain information on the instrument itself and the rationale behind its use (the expanded chart entries).

DEVIATION IQ (DIQ) - An age-based index of general mental ability. It is based on the deviation or difference between a person's obtained score and the score that is average for persons of that chronological age. DIQ is the commonly used "IQ" score, although it is technically different in derivation from an actual IQ (non-deviation value).

DIAGNOSTIC TEST - A test used to locate an individual's specific areas of weakness or strength, to determine the nature of his weakness or deficiencies, and to suggest their cause whenever possible.

DIFFICULTY VALUE - An index which indicates the percent of some specified group, such as students of a given age or grade, who answered a test item correctly. It can be used with an item analysis report as reference data.

DISCRIMINATION INDEX - An index which indicates the discriminating power of a test item. The most commonly used index is derived from the number passing the item in the highest 27 percent of the group (on total score) and the number passing in the lowest 27 percent.

DISTRACTOR - Any incorrect choice (option) in a test item.

ERROR OF MEASUREMENT - As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due

to errors of measurement. The larger the standard error of measurement, the less reliable the score. The standard error of measurement is an amount such that in about two-thirds of the cases, the obtained score would not differ by more than one standard error of measurement from the true score.

EXPECTANCY TABLE ("expected" achievement) - A term with two common usages, related but with some differences:

(1) A table or other device for showing the relation between scores on a predictive test and some related outcome. The outcome, or criterion status, for individuals at each level of predictive score may be expressed as (a) an average on the outcome variable, (b) the percent of cases at successive levels, or (c) the probability of reaching given performance levels. Such tables are commonly used in making predictions of educational or job success.

(2) A table or chart providing for an interpretation of a student's obtained score on an achievement test with the score which would be "expected" for those at his grade level and with his level of scholastic aptitude. Such "expectancies" are based upon actual data from administration of the specified achievement and scholastic aptitude tests to the same student population. The term "anticipated" is also used to denote achievement as differentiated by level of "intellectual status."

FACE VALIDITY - Refers to the acceptability of the test. A test has face validity when it appears to measure the variable to be tested.

FORMATIVE - Evaluations used to determine the degree of mastery of a given goal or learning task and to pinpoint the part of the goal or task not attained or mastered. Evaluations and tests used for formative purposes are part of a process; they help to form subsequent steps in the process. This differs from summative usages (which sum up) that come at the end of a course and serve mostly for grading or goals met types of evaluation.

FREQUENCY DISTRIBUTION - A tabulation of scores from high to low showing the number of individuals that obtained each score or fell in each score interval.

GRADE EQUIVALENT (GE) - The grade level for which a given score is the real or estimated average. Grade-equivalent interpretation, most appropriate for elementary level achievement tests, expresses obtained scores in terms of grade and month of grade, assuming a 10-month school year (e.g., 5.7). Since such tests are usually standardized at only one (or two) point(s) within each grade, grade equivalents between points for which there are data-based scores must be "estimated" by interpolation.

INSTRUMENT - The general class name of measures, tests, and assessment techniques.

INTELLIGENCE - The ability to perceive and understand relationships, such as logical, spatial, verbal, numerical, and recall of associated meanings.

INTELLIGENCE QUOTIENT (IQ) - Originally, an index of brightness expressed as the ratio of a person's mental age to his chronological age (MA/CA), multiplied by 100 to eliminate the decimal. The quotient IQ has been gradually replaced by the deviation IQ concept (DIQ).

ITEM ANALYSIS - Anyone of several methods used in test construction to determine how well a given test item discriminates among individuals differing in some characteristic.

MASTERY TEST - A test designed to determine whether a pupil has mastered a given unit of instruction or skill.

MEAN - The sum of a test of scores divided by the number of scores.

MEDIAN - The middle score in a set of ranked scores. It is the point above or below which an equal number of ranked scores lie. It corresponds to the 50th percentile.

MODE - The score or value that occurs most frequently in a distribution.

NORMAL DISTRIBUTION - A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. In such a normal distribution, scores or measure are distributed symmetrically about the mean. Cases are concentrated near the mean and decrease in frequency the farther one departs from the mean.

NORMAL DISTRIBUTION CURVE - A derived curve based on the assumption that variations from the mean are by chance. It is bell-shaped in form and adopted as true because of its repeated recurrence in the frequency distributions of sets of measurements of human characteristics in psychology and education. It has many useful mathematical properties. In a normal distribution curve, scores are distributed symmetrically about the mean, as many cases at various equal distances above the mean as below the mean, and with cases concentrated near the average and decreasing in frequency the farther one departs from it.

NORM-REFERENCED (Measurement, tests) - The construction and scoring of psychological tests which yield scores that are interpreted on scales which are derived from the statistical relationships of groups of scores by other individuals, selected according to sampling strategy. This specially selected group is usually designated as one of the tests' norming group.

NORMS - Statistics that supply a frame reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization

group for the test. Since they represent average or typical performance, they should not be regarded as standards or universally desirable levels of attainment.

OBJECTIVE (Instruction, et al) - A conceptual tool used in means and ends analysis with respect to formation of learning opportunities, and instructional and assessment strategies. Mediates a learning outcome with the choice of means expected to bring about a particular learning in the individual.

OPPORTUNITY (Learning) - Any occasion upon which an individual might learn something as the result of being present and sentient for the duration of the occasion.

OUTCOMES (Learning) - A descriptive statement of what an individual will know, be capable of doing, or probably disposed to do, as a result of having learning something.

PERCENTILE (P) - A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. A score coinciding with the 60th percentile (P60) is regarded as equaling or surpassing that of 60 percent of the persons in the group, with 40 percent of the performances exceeding this score. A student's converted score indicating his relative position in a specific group may be expressed as his percentile rank.

PERCENTILE BAND - An interpretation of a test score which takes account of the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above it.

PERCENTILE RANK (PR) - The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

PERFORMANCE TEST - Commonly used to designate a test in which the subject responds by overt action, rather than by making a verbal or written response. The test intends to measure actual accomplishment rather than potential ability or aptitude.

POWER TEST - A test which is designed to sample the range of an examinee's capacity in particular skills or abilities and which places minimal emphasis on time limits.

RANDOM SAMPLE - A sample of the members of some total population drawn in such a way that every member of the population has an equal chance of being included. It is done in such a way that precludes the operation of bias or selection.

RAW SCORE - The first quantitive result obtained in scoring a test. Usually the number of questions answered correctly or the number right minus some fraction of the number of wrong answers.

READINESS TEST - A test of ability to engage in a new type of specific learning. Level of maturity, previous experience, and mental and emotional set are important factors in readiness.

REGRESSION EFFECT - Tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure.

RELIABILITY - The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement.

REPRESENTATIVE SAMPLE - A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation. In an achievement test norm sample, such significant aspects might be the proportion of cases of each sex, from various types of schools, different geographical areas, several socioeconomic levels.

SCHOLASTIC APTITUDE - A combination of native and acquired abilities that is needed to do school work.

SKEWNESS - The degree to which a unimodal (one-peaked) curve departs from symmetry.

STANDARD DEVIATION (S.D.) - A calculated value which expresses the way a group of scores spreads out, or is clustered around the mean. In a normal distribution, plus and minus one standard deviation from the mean sets off approximately 2/3 of the population. The smaller the S.D., the more homogeneous the group being examined.

STANDARD SCORE - A general term referring to a score distribution which has been "standardized" to a specific mean and a specific standard deviation. This allows raw scores to be converted to standard scores for convenience, comparability, and ease of interpretation.

STANINES - A unit that divides the norm population into nine groups. Except for Stanines 1 and 9, the groups are spaced in half-sigma units, with the mean at Stanine 5.

Stanine	1	2	3	4	5	6	7	8	9
% in Stanine	4	7	12	17	20	17	12	7	4

STRATIFIED SAMPLE - A sample in which cases are selected by the use of

certain controls, such as geographical region, community size, grade, age, sex, etc.

SUMMATIVE - Evaluations used as a general assessment of the degree to which goals or learning tasks have been attained over an entire program or course, or some substantial part of it.

T-SCORE - A derived score based upon the equivalence of percentile values to standard scores, thus avoiding the effects of skewed distributions, and usually having a mean equated to 50 and a standard deviation equated to 10.

VALIDITY - A test is valid to the extent that we know what it measures or predicts. There are five types of validity:

1. CONTENT VALIDITY - refers to how well the content of the test samples of the subject matter or situation about which conclusions are to be drawn. Content validity is especially important in an achievement test.
2. CONSTRUCT VALIDITY - Concerns the psychological qualities of test measures. By both logical and empirical methods, the theory underlying the test is validated.
3. CONCURRENT VALIDITY - Refers to how well test scores match measures of contemporary criterion performance.
4. ITEM STRUCTURE - Includes (1) corroborative evidence from ITEM ANALYSIS supporting the other characteristics of the test; i.e., the inter-relationships between items, between items and scores, and between items and criteria are factors that contribute to the content validity of the test, and (2) item composition. For graphic items, this emphasizes perceptual clarity and related format functions. For verbal items, it emphasizes conceptual clarity in the expression of items. And for both graphic and verbal items it emphasizes the functions of distractors.
5. PREDICTIVE VALIDITY - Relates to how well predictions made from the test are confirmed by data collected at a later time.

VARIABILITY - The spread or disperson of scores, usually indicated by quartile deviations, range of 90-10 percentile scores, etc.